Organization of Information for the Web using Hierarchical Fuzzy Clustering Algorithm based on Co-Occurrence Networks

Faraz Zaidi, Guy Melançon

Organization of Information for the Web using Hierarchical Fuzzy Clustering Algorithm based on Co-Occurrence Networks

Faraz Zaidi and Guy Melançon
CNRS UMR 5800 LaBRI & INRIA Bordeaux Sud-Ouest
Bordeaux, France
{*faraz.zaidi, guy.melancon*}*@labri.fr*

***Abstract*—In this paper, we present a Hierarchical Fuzzy Clustering algorithm which uses domain knowledge to automatically determine the number of clusters and their initial values. The algorithm is applied on a collection of web pages and the results are compared with existing algorithms in the literature.**

***Keywords*-Web Mining; Fuzzy and Hierarchical Clustering; Information Retrieval**

I. Introduction

Searching information on the web is a common task where the goal is to return a number of web pages corresponding to the searched keywords entered by the user. Typically, document clustering techniques consider *documents* as entities and calculate a similarity between documents based on the *keywords* appearing in these documents. This similarity is further used to group similar documents together to obtain clusters. Intuitively, an alternative approach is to consider *keywords* as entities related to each other if they appear in a single document. Clustering keywords can group *Themes* together such that when a keyword is searched, it's theme can be identified from the cluster it belongs to. Obviously documents can be associated to these *themes* and returned as search result to the user. There are a number of issues that need to be addressed before selecting the right clustering methodology which are discussed below.

An important issue is whether the clustering algorithm should generate Hard clusters or Soft clusters i.e. documents (or web pages) should belong to unique clusters or may be associated to more than one cluster. Researchers have shown that very often, a web page can belong to more than one category and thus it is more useful to use Soft or Fuzzy clustering algorithms [1]. Another important decision is to choose between Hierarchical or Flat clustering. Naturally, information around us is organized in a hierarchical manner. Again this claim is supported by many researchers that tend to organize documents and web pages in hierarchies [6].

Having said that we need hierarchical and fuzzy clustering approaches to cluster web pages, we would like to point out some other requirements induced by the domain, that is, the web. In terms of hierarchical clustering, generating a hierarchy of high depth is not suitable, as the famous Three-Click Rule [13] suggests, users tend to abandon a site if they don't find their required information with-in three clicks. Another requirement from the fuzzy perspective is that once we have calculated the degree of similarity of a web page to various clusters, we need to find a threshold which assures that only the relevant pages are grouped together. Finally the classical problem of deciding the number of clusters to be generated, in case of a hierarchical algorithm, the number of clusters generated at each level are also important.

In this paper, we take a different approach to solve all these issues by looking at the co-occurrence network of the keywords of the web pages. A co-occurrence network is a graph where the nodes are represented by *keywords* and edges between keywords imply that they appear together at least once in a web page. These networks have some interesting properties that can be used to devise heuristics which can eventually help us resolve issues described earlier. Based on these properties, we propose a new Hierarchical Fuzzy Clustering Algorithm based on Co-Occurrence Networks (HFC-CN) where we claim that the algorithm does not require any parameter as input. The performance of the proposed algorithm is compared with other existing algorithms and the results are quite satisfactory.

For experimentation, we have used three different data sets. These data sets are a collection of web pages found on Wikipedia encyclopedia. These web pages were returned as a search result when *Jaguar*, *CAC40* and *Hepburn* were launched as a query on the Exalead (http://www.exalead.com/search/wikipedia/) search engine. In each case, the top 50 results were considered and *keywords* from these web pages were extracted by the Exalead Search Engine.

II. Related Work

There are a number of document clustering algorithms present in the literature. An exhaustive overview of these algorithms remains out of the scope of this paper. Our focus in this paper is information retrieval, which in the current context means that a user has one or more than one key words and the goal is to search for documents containing those keywords within a collection of documents.

As described previously, in this paper, we present a new approach to document clustering which is based on clustering keywords. These clusters are then used to regroup

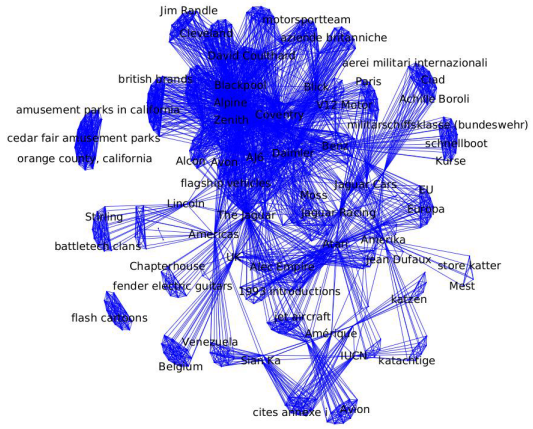

Figure 1. Co-Occurrence Network for *Jaguar* keyword where disconnected components can be easily identified as forming a clique.

documents, where the degree of similarity of a document is calculated to a cluster of keywords. As a result, a document can belong to more than one cluster. The idea of clustering words is not a new idea. [5], [7] have used word clustering to reduce dimensions of a document and then cluster documents eventually. Our approach presents a solution to the information retrieval problem and has clearly different objectives as presented in [5].

There are a number of Hierarchical Fuzzy Clustering algorithms proposed in the literature. [4] provides a good overview of these algorithms. A general drawback for these algorithms is the number of input parameters required to determine the number of clusters, the initial centroids and the hierarchy cut. Our approach differs from these algorithms as it does not require any parameter for execution. We have compared our results with two such algorithms that require minimal number of parameters and show the results in section IV. The Fuzzy Agglomerative Hierarchical Clustering (FAHC) Algorithm introduced by [10] and the Hierarchical-Hyperspherical Divisive Fuzzy C-Means (H2D-FCM) Algorithm [4]. Readers are referred to the respective articles for details of these algorithms.

III. Proposed Algorithm

The co-occurrence networks constructed from web pages have two interesting properties. First, the node degree distribution in these networks is not random, or they do not follow a Gaussian Distribution (see Figure 2). They have a long tail like structure representing the scale free degree distribution [2] suggesting that there are nodes in the network with very high node degree or connectivity. Another important property is that every node in this network belongs to a clique. This property is inherited by construction as all the keywords extracted from a single document are connected to each other by an edge in the co-occurrence network, thus forming a clique as shown in Figure 1. We exploit both these properties to determine the number of clusters as well as their centroids where the details are explained below.

A. Detection of Number of Clusters and Centroids

In this section, we describe a method to determine the number of clusters inspired by an earlier work of the authors [11]. The method is based on a decomposition technique which exploits the fact that nodes having high degree are responsible for keeping large size networks as a single connected component. This fact can be used to identify, what we call themes or subjects around which the different web pages are organized.

The goal is to identify the keywords that have a relatively high connectivity to other nodes but are not present in all the documents. We define a Min$_d$-Degree Induced Subgraph (Min$_d$-DIS) which is an induced subgraph of nodes having degree at least d in graph G. The variable d can have values from 0 to the maximum node degree possible for a network. We construct Min$_d$-DIS for $d = \{0, 1, \cdots, MaxDeg\}$ to obtain a set of graphs $(G_0, G_1, \cdots, G_{MaxDeg})$. Consider Figure 3 and Figure 4 where the Min$_{80}$-DIS and Min$_{50}$-DIS graphs are laid out from the example *Jaguar*. The Min$_{80}$-DIS contains the nodes that have at least degree 80 in the co-occurrence network. Looking at the connectivity of the nodes, all these nodes form a clique as they are connected to each other, representing that they all appear together in documents, lets call these nodes as the *core* nodes of the network. Moving from high values of node degree, we eventually come across nodes that are not connected to the core nodes in the Min$_{80}$-DIS graph. Figure 4 is such an example where Min$_{50}$-DIS contain nodes that are not connected to every other node. Here we clearly see that if the core nodes (found in Min$_{80}$-DIS) are removed from this graph, we will be left with three connected components, the component with the nodes labeled *Ferrari* and *Sports Cars*, the *Jaguar Cat* and *Atari*. By repetitive application of this method, we can identify these components as the basic themes of the collection of pages.

Once we have identified these basic themes, we need to find the centroids for these themes. An additional information which can be used to devise the centroids for these themes comes from the fact that by construction, in the co-occurrence network, every node belongs to a clique (as all the words belonging to a document are connected to each other in the co-occurrence network). So if we look at Figure 4, the word *jaguar cat*, *Atari* and *Ferrari, Sports Cars* must belong to different cliques. Moreover, since by construction of the Min$_{50}$-DIS, these nodes have a node degree of at least 50. This suggests that these nodes are important in this collection of documents. To find the collection of documents to which these themes belong to, we simply group the nodes that are at distance 1 from each of these nodes in the entire network. In this way, we identify a group of documents belonging to a theme which can be

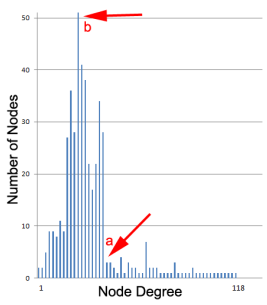

Figure 2. Degree Distribution of Jaguar Co-Occurrence Network.

used to calculate the centroid of the cluster for this theme. The algorithm to detect the number of clusters and their centroids is listed below:

$Input\ G(V, E)$
$var = MaxDeg$
while $var \geq cutoff$ **do**
  $G' = calculate(Min_{var} - DIS)$
  if $isNotClique(nodes(G'))$ **then**
    $Call\ Procedure\ IdentifyThemes(G')$
  end if
  $var = var - 1$
end while
$Procedure\ IdentifyThemes(G')$
for all i such that $i\ not\ connected\ to\ all\ nodes\ in\ G'$ **do**
  $Find\ Nodes\ connected\ to\ i\ at\ distance\ 1\ in\ G$
  $Group\ Nodes\ as\ a\ Centroid$
end for

Algorithm: Detection of Number of Clusters and their Centroids.

The above algorithm requires a parameter $cutoff$ as input which represents the value up to which the Min$_d$-DIS is calculated. To determine this value, we use a heuristic proposed by [12] to determine the high degree nodes in a scale free network. Looking at Figure 2, it is quite clear that there are nodes that dominate the number of connections by having a high node degree. Semantically, it is quite obvious, if we search for the word *Jaguar* on the web, all the pages returned will surely have this word and thus would have a very high degree as compared to the other words appearing in this collection of web pages.

To find out these high degree nodes, we calculate the slope of every two consecutive points of the degree distribution . At point *a* in Figure 2, the slope becomes equal to zero. The heuristic suggests that as the slope becomes zero or close to zero (values of -1 or -2) the point can be considered as the cutoff point where the nodes lying after this point represents the nodes that have relatively high node degree as compared to other nodes in the network. Since our goal is to generate a hierarchical clustering, we need to generate

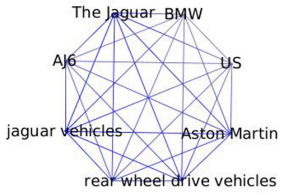

Figure 3. Min$_{80}$-DIS for the Jaguar example

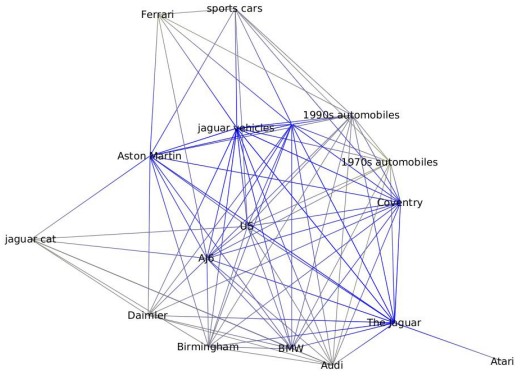

Figure 4. Min$_{50}$-DIS for the Jaguar example

different *cutoff* points to incorporate the multilevel structure. Another point that stands out in the degree distribution of these co-occurrence networks is at some value for node degree, the number of nodes attain a maximum number, as pointed by *b* in Figure 2. Since our goal is to generate a hierarchy of up to three levels, the second cutoff is considered to be the arithmetic mean of point *a* and *b* where as the third cutoff is the point *b*.

B. HFC-CN Algorithm

Once we have calculated the cluster centroids, the remaining algorithm to generate a fuzzy clustering is quite simple. First we cluster all the remaining nodes in the co-occurrence network by assigning them to one of the centroids. As a result, we generate a hard and partitional clustering for the network. We then calculate for each document in the collection its degree or relevance to these centroids giving us a fuzzy clustering where a document can belong to more than one centroid. To generate a hierarchical clustering, we run the algorithm for different values of $cutoff$ where at each level, only the nodes belonging to a cluster are considered as compared to the whole network.

The resulting algorithm gives us a divisive hierarchical fuzzy clustering for the document collection. Each document has an associated degree of relevance to the to other cluster centroid which in some cases can be 0 as well. The results of the quality of clustering produced are tabulated in Table I.

IV. Experimentation and Results

To measure the validity of the clustering produced by the proposed algorithm as compared to the FAHC and H2D-

Partition Coefficient	Clustering Algorithms		
	FAHC	H2D-FCM	HFC-CN
Jaguar	0.415	0.357	0.349
Hepburn	0.385	0.317	0.357
CAC40	0.279	0.252	0.279
Partition Efficiency			
Jaguar	0.566	0.736	0.607
Hepburn	0.438	0.479	0.463
CAC40	0.456	0.504	0.495

Table I
COMPARING THE RESULTS OF CLUSTERING ALGORITHMS USING PARTITION COEFFICIENT AND PARTITION EFFICIENCY.

FCM algorithms, we use two validity indices used in the fuzzy environment. The Partition Co-efficient(PC) [8] and Partition Efficiency(PE) [3]. Both these methods are based on only the membership values[9] of an artifact to various clusters. The PC index indicates the average relative amount of membership sharing done between pairs of fuzzy subsets. The values range in $[1/c, 1]$ where c is the number of clusters. The PE index is a scalar measure of the amount of fuzziness in a given fuzzy clustering where the values range in $[0, logc]$. In both these cases low values indicate high clustering quality. To handle the hierarchical clustering, for each level we compute these validity indices and then we take the average over the different hierarchical levels. Recall that we have forced the algorithms to produce a hierarchy of at most 3 levels.

Table I shows the results obtained by the HFC-CN algorithm as compared to the other two clustering algorithms using Partition Coefficient and Partition Efficiency respectively. It can easily be concluded that the algorithm performs as well as the other algorithms and determines correctly the number of clusters as well as the cluster centroids.

An important feature of the proposed algorithm as compared to other algorithms is the way the initial centroids are calculated. The other algorithms use only one single document as a centroid either chosen randomly, or based on the dissimilarity of existing centroids and a new document. The proposed method identifies important keywords and then calculates the initial cluster centroids based on a number of documents containing those keywords. Moreover, since the clustering is based on similarity of words as compared to similarity of documents, the topics that are similar based on some theme are grouped together and we only calculate the similarity of documents to the set of words that are clustered together. This seems to work well as semantically when we look at the clusters produced by the clustering algorithm, they are indeed coherent.

V. CONCLUSION

In this paper, we have presented a divisive fuzzy clustering algorithm for documents using graph theoretical concepts on the co-occurrence network of keywords obtained from a collection of web pages. We have addressed the well know problems of the detection of number of clusters, the initial centroids and the depth of hierarchy to be generated in the context of information retrieval and web pages. Comparative results show that the proposed method performs well as two other well known algorithms used to produce Hierarchical Fuzzy Clustering for documents.

REFERENCES

[1] D. Arotaritei and S. Mitra. Web mining: a survey in the fuzzy framework. *Fuzzy Sets and Systems*, 148(1):5 – 19, 2004. Web Mining Using Soft Computing.

[2] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[3] J. C. Bezdek. Cluster validity with fuzzy sets. *Cybernetics and Systems*, 3(3):58–73, 1973.

[4] G. Bordogna and G. Pasi. Hierarchical-hyperspherical divisive fuzzy c-means (h2d-fcm) clustering for information retrieval. In *WI-IAT*, pages 614–621, 2009.

[5] I. S. Dhillon, S. Mallela, and R. Kumar. Enhanced word clustering for hierarchical text classification. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 191–200, New York, NY, USA, 2002. ACM.

[6] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. Technical report, Departement of Computer Science and Engineering, University of Minnesota, 2000.

[7] W. C. Tjhi and L. H. Chen. Possibilistic fuzzy co-clustering of large document collections. *Pattern Recognition*, 40(12):3452–3466, Dec. 2007.

[8] E. Trauwaert. On the meaning of dunn's partition coefficient for fuzzy clusters. *Fuzzy Sets Syst.*, 25(2):217–242, 1988.

[9] W. Wang and Y. Zhang. On fuzzy cluster validity indices. *Fuzzy Sets Syst.*, 158(19):2095–2117, 2007.

[10] Y.-C. C. C.-H. L. Yih-Jen Horng, Shyi-Ming Chen. A new method for fuzzy information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques. *Fuzzy Systems, IEEE Transactions on*, 13(2):216 – 228, April 2005.

[11] F. Zaidi and G. Melançon. Identifying the Presence of Communities in Complex Networks Through Topological Decomposition and Component Densities. In *EGC 2010*. 163-174, 2010.

[12] F. Zaidi, A. Sallaberry, and G. Melançon. Revealing hidden community structures and identifying bridges in complex networks: An application to analyzing contents of web pages for browsing. In *WI-IAT '09*, pages 198–205, 2009.

[13] J. Zeldman. *Taking Your Talent to the Web: A Guide for the Transitioning Designer*. New Riders Publishing, Thousand Oaks, CA, USA, 2001.